Estimating Local Intrinsic Dimensionality

L. Amsaleg[1] O. Chelly[2] T. Furon[1]
S. Girard[3] M. E. Houle[2] K. Kawarabayashi[2] M. Nett[24]

[1]Equipe TEXMEX, INRIA/IRISA Rennes, France

[2]National Institute of Informatics, Chiyoda-ku, Tokyo, Japan

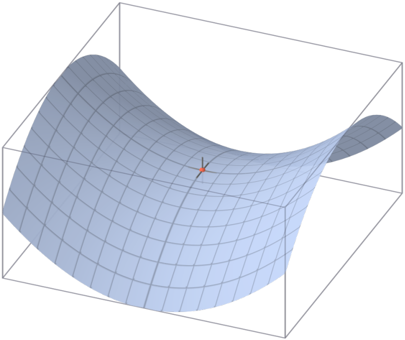
[3]Equipe MISTIS, INRIA Grenoble Rhône-Alpes, France

[4]Google Japan, Minato-ku, Tokyo, Japan

August 3, 2015

Table of contents

What is intrinsic dimensionality?



Definition: Intrinsic Dimensionality

Intrinsic dimensionality can be described as the minimum number of coordinates required to locate a point in the space.

Current applications of ID

Analysis of search indices (Expansion Dimension)

- Navigating Nets [Krauthgamer & Lee (2005)]
- Cover Tree [Beygelzimer, Kakade, Langford (2006)]
- Rank Cover Tree [Houle & Nett (2013)]

Analysis of projection, Outlier detection (Expansion Dimension)

- LOF-based PINN outlier detection [de Vries, Chawla, Houle (2010)]

Projection and dimensional reduction

- Principal component analysis [Pearson (1901)]

Potential applications of ID

Prediction

- estimate the target dimension for dimensionality reduction.
- predict the difficulty of data sets/subsets/points.

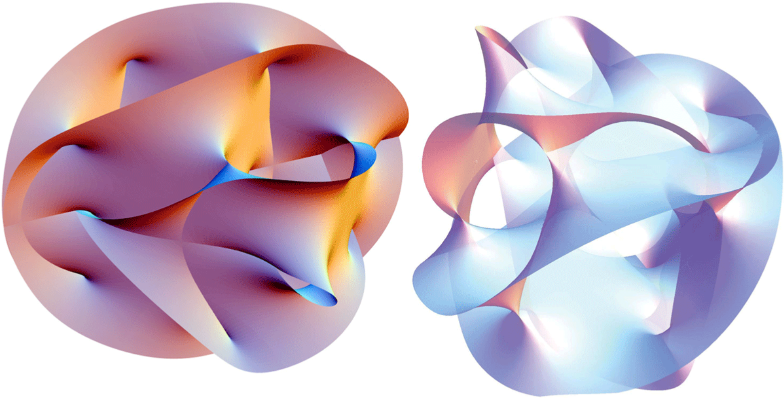
Processing

- redesign machine learning algorithms by adapting them to the local variation of ID (queries in points with low ID are more trustworthy).

Evaluation

- explain the behavior of algorithms and evaluate them more fairly by accounting for the disparity in ID.

Global models of ID vs. Local models of ID

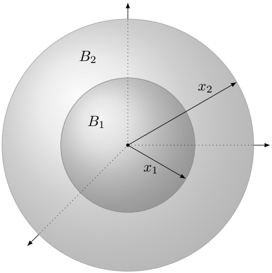


Global models of ID vs. Local models of ID

Properties of global and local models

	Global models	Local models
measure	the dimensinality of the whole dataset.	the dimensionality in the neighborhood of a point.
require	a set of data points.	a set of distances to the nearest neighbors.
examples	Topological models (PCA), fractal models (Hausdorff Dimension, Correlation Dimension), Graph-based models, etc.	Expansion models (ED, LID).

Expansion models



Expansion-based methods estimate the intrinsic dimension by comparing the expansion in distance and the associated expansion in volume (number of points).
Examples: Expansion Dimension (ED), Minimum Neighborhood Distance (MiND), Local Intrinsic Dimension (LID).

How to measure the dimension?

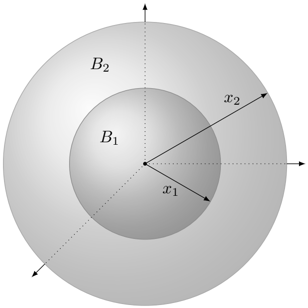
Dimensional query

In an L_p-norm space of dimension m, if V is a measure of volume then :

$$\frac{V(B_2)}{V(B_1)} = \left(\frac{x_2}{x_1}\right)^m$$

The representational dimension m can be obtained by :

$$m = \frac{\ln V(B_2) - \ln V(B_1)}{\ln x_2 - \ln x_1}$$

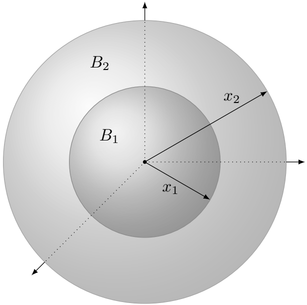


Expansion Dimension

Expansion dimension

If volume is measured in terms of number of points that are captured, then

$$\mathrm{ED}(q, x_1, x_2) = \frac{\ln k_2 - \ln k_1}{\ln x_2 - \ln x_1}$$



Distance as a continuous random variable

Main assumption

Data can be seen as a sample generated from a set of continuous random variables.

Continuous random distance variable

Let **X** be an absolutely continuous random distance variable that represents the distance from a reference point q with

- probability density function f_X
- cumulative distribution function $F_X = \int f_X(t)dt$
- $(x_i)_{1 \leq i < n}$ an ordered sample of X

Intrinsic dimensional query

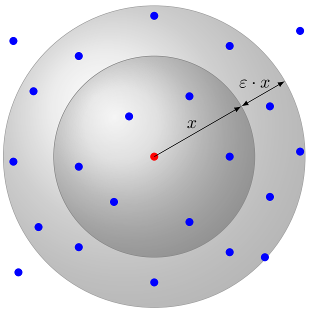
Definition

When $F_X(x) > 0$, the continuous ID of X at distance x is defined as

$$\mathrm{IntrDim}_{F_X}(x) = \lim_{\varepsilon \to 0^+} \left(\frac{\ln F_X((1+\varepsilon)x) - \ln F_X(x)}{\ln(1+\varepsilon)} \right)$$

Remarks

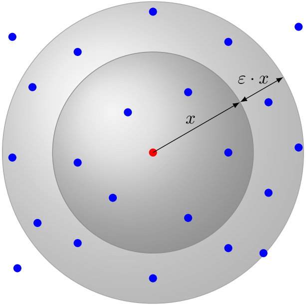
- The volume is measured in terms of expected number of points.
- X depends on reference point q.



Intrinsic dimensional query

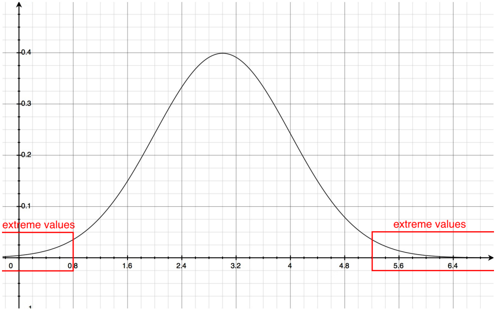
By applying l'Hôpital's rule:

$$\mathrm{ID}_{F_X}(x) = \frac{x \cdot f_X(x)}{F_X(x)}$$



Disaster prevention

Introduction to EVT



Analogy between central and extreme values

- Central values → Central limit theorem → Normal distribution
- Extreme values → Pickands-Balkema-de Haan theorem → Generalized Pareto Distribution

Generalized Pareto Distribution

Cases of the Generalized Pareto Distribution

- $\xi = 0 \rightarrow$ Gumbel family
- $\xi > 0 \rightarrow$ Fréchet family
- $\xi < 0 \rightarrow$ Weibull family (The distribution is upper bounded)

- Distance distributions are lower bounded.
- We can model distances under a threshold w using Weibull distribution.

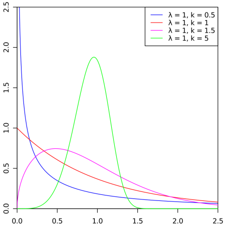


Figure: pdf of the Weibull distribution

Modeling the tail of a distance distribution

Modeling threshold excesses

As a certain threshold w approaches 0, the excess $Y = w - X$ follows a GPD with parameters ξ and σ:

$$\Pr[Y \leq y | Y < w] \approx 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}}$$

Heuristic

As $w \to 0$, the distribution of X restricted to the tail $[0, w)$ converges to that of a random distance variable X^*

$$F_{X,w}(x) \approx F_{X^*,w}(x) = \left(\frac{x}{w}\right)^{-\frac{1}{\xi}}$$

Modeling the tail of a distance distribution

Modeling threshold excesses

$$\mathrm{ID}_{F_X}(0) \approx \lim_{x\to 0} \mathrm{ID}_{F_{X^*}}(x) = \lim_{x\to 0} \frac{x \cdot f_{X^*,w}(x)}{F_{X^*,w}(x)} = -\frac{1}{\xi}$$

Remarks

- This is an approximation since it holds for the limit distribution as $w \to 0$.
- We can use classical estimation methods to estimate the parameter ξ.
- Note: In EVT, $-1/\xi$ is called the index of the distribution.

Usual statistical methods

Maximum likelihood estimator

$$\widehat{\mathrm{ID}}_{F_{X^*}}(0) = -\left(\frac{1}{n}\sum_{i=1}^{n} \ln \frac{x_i}{x_n}\right)^{-1}$$

Method of moments estimator

$$\widehat{\mathrm{ID}}_{F_{X^*}}(0) = -k\frac{\hat{\mu}_k}{\hat{\mu}_k - x_n^k} \quad \text{where} \quad \hat{\mu}_k = \sum_{i=1}^{n} x_i^k$$

Probability weighted moments estimator

$$\widehat{\mathrm{ID}}_{F_{X^*}}(0) = \frac{\hat{\nu}_k}{w - (k+1)\hat{\nu}_k} \quad \text{where} \quad \hat{\nu}_k = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{i - 0.35}{n}\right)^k x_i$$

Regularly Varying Functions Estimator

Regularly Varying Functions Estimator

$$\widehat{\mathrm{ID}}_{F_{X^*}}(0) = \hat{\kappa} = \frac{\sum_{j=1}^{J} \alpha_j \ln\left[\hat{F}_X((1+\tau_j\delta_n)x_n)/\hat{F}_X(x_n)\right]}{\sum_{j=1}^{J} \alpha_j \ln(1+\tau_j\delta_n)}$$

under the assumption that $\delta_n \to 0^+$ as $n \to \infty$
where : $(\alpha_j)_{1\leq j<J}$ and $(\tau_j)_{1\leq j<J}$ are sequences.

LID estimation in artificial distances

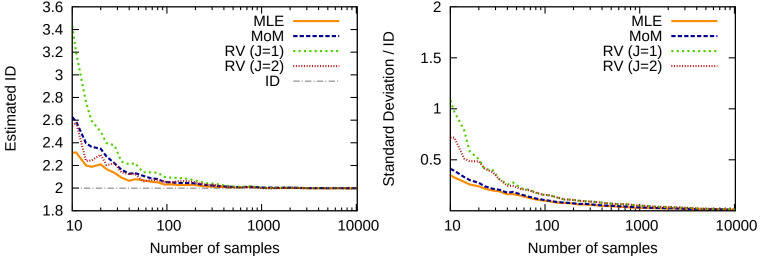


Figure: Comparison of ID estimates (ID=2)

LID estimation in artificial distances

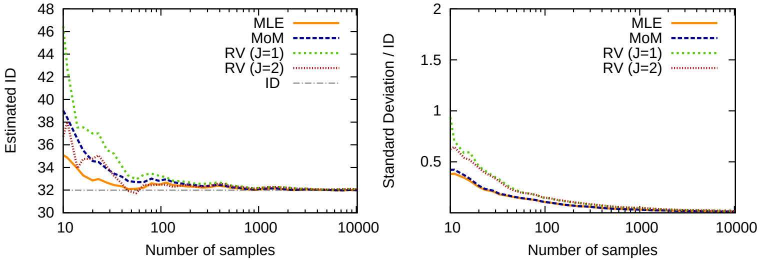


Figure: Comparison of ID estimates (ID=32)

ID estimation in artificial datasets

Datasets

manifold	d	D	description
1	10	11	Uniformly sampled sphere
2	3	5	Affine space
3	4	6	Concentrated figure confusable with a 3d one
4	4	8	Non-linear manifold
5	2	3	2-d Helix
6	6	36	Non-linear manifold
7	2	3	Swiss-Roll

ID estimation in artificial datasets

Datasets

manifold	d	D	description
8	12	72	Non-linear manifold
9	20	20	Affine space
10a	10	11	Uniformly sampled hypercube
10b	17	18	Uniformly sampled hypercube
10c	24	25	Uniformly sampled hypercube
11	2	3	Möbius band 10-times twisted
12	20	20	Isotropic multivariate Gaussian
13	1	13	Curve

What is the Intrinsic Dimension?

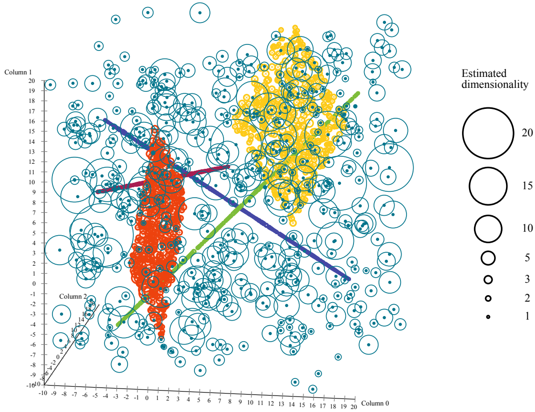


Figure: Data structures are detected by ID

ID estimation in artificial datasets

Conclusions

- Local estimators tend to over-estimate the dimensionality of non-linear manifolds, and to under-estimate that of linear manifolds.
- For nonlinear manifolds, global estimators have difficulty in identifying the intrinsic dimension.
- The higher the sampling rate, the lower the bias.
- Global methods are very affected by noise, while local methods are more resistant.

Values of LID in real datasets

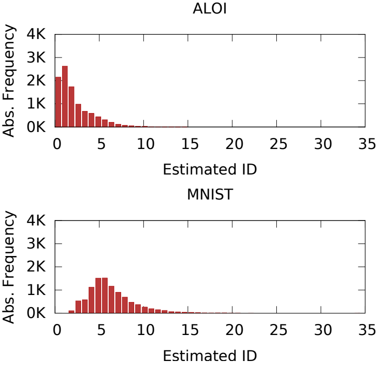


Figure: LID MLE estimates in ALOI and MNIST

Thank you for your attention!